University of Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: Master of Science in Data Science	Spring semester, 20..23.. Open / Restricted access
Writer: Kevin Mekhaphan Nguyen	Kevin M. Nguyen ..... (Writer's signature)
Faculty supervisor: Alvaro Fernandez-Quilez External supervisor(s):	
Thesis title: Uncertainty quantification in prostate segmentation	
Credits (ECTS): 30	
Key words: Conformal prediction, CNN, nnU-net, prostate cancer, segmentation, deep learning	Pages: 39 …… + enclosure: …… Stavanger, 15/12/2023.. Date/year

Front page for master thesis
Faculty of Science and Technology
Decision made by the Dean October 30th 2009

University of Stavanger

Faculty of Science and Technology
Department of Electrical Engineering and Computer Science

Uncertainty quantification in prostate segmentation

Master's Thesis in Computer Science
by

Kevin Mekhaphan Nguyen

Internal Supervisors

Alvaro Fernandez-Quilez

December 15, 2023

Abstract

Prostate cancer, a significant global health challenge, necessitates innovative diagnostic solutions. Despite the invaluable role of Magnetic Resonance Imaging (MRI), challenges persist in analysis due to time-intensive tasks and inter-reader variability. Accurate prostate segmentation is critical for diagnosis, influencing clinical decisions and further testing choices.

Traditionally, Convolutional Neural Networks (CNNs) have been employed for automated segmentation tasks, but the manual assessment of segmentation quality remains a crucial bottleneck. This research shifts the paradigm by exploring statistical approaches, specifically focusing on Conformal Prediction (CP), to evaluate the quality of prostate segmentation. Clinically relevant metrics, including Dice Score, relative volume difference, efficiency, and validity, are employed for quantitative assessment and comparison. The conformal classifier demonstrates robustness across diverse datasets. Nearest-Neighbor interpolation ensures image resizing uniformity, and patient-centric data splitting with Region of Interest (ROI) extraction enhances the model's focus.

The work we present is an innovative approach in prostate cancer segmentation using conformal prediction. It focuses on quantifying uncertainties in segmentation and evaluates segmentation quality through the Dice Score and RVD metrics. The study stands out for its high validity and efficiency, achieving percentages ranging from 94.24% to 99.34% on external datasets. This approach significantly enhances the diagnostic accuracy in prostate cancer detection via MRI analysis, showcasing the potential of integrating conformal classification in medical imaging to improve precision in clinical diagnostics.

This research advances prostate cancer diagnosis methodologies, emphasizing the novel application of conformal prediction for quantifying the segmentation obtained by other deep learning models. The findings underscore the importance of precise segmentation quality assessment, emphasizing the significance of specific metrics in evaluating the proposed statistical approach for quality control in prostate cancer diagnosis.

Acknowledgements

I would like to thank my supervisor, Alvaro Fernandez-Quilez, for the help and guidance on the writing and completion of the thesis.

Contents

Chapter 1

Introduction

1.1 Motivation

Prostate cancer, ranking as the fifth leading cause of death globally, continues to present formidable challenges in its diagnosis and subsequent treatment [1]. Magnetic Resonance Imaging (MRI) has proven to be a valuable asset in aiding the identification and evaluation of prostate cancer [2]. However, the analysis of MRI data for prostate cancer diagnosis is complex, involving significant time investment and facing challenges related to inter-reader variability among radiologists [3].

Accurate segmentation of the prostate is pivotal in the diagnosis of prostate cancer, influencing critical clinical measures and subsequent decisions for additional testing and treatment. While Convolutional Neural Networks (CNNs) have been widely employed for automated prostate segmentation, the manual assessment of segmentation quality remains a bottleneck, introducing subjectivity and potential inconsistencies [4].

Motivated by the vital need for a standardized and efficient approach, this study delves into the exploration and testing of conformal prediction techniques for quantifying the uncertainties of prostate segmentations done by other deep learning models [5]. Specifically, we apply class-conditional conformal classification to the predicted probabilities generated by previous trained deep learning models. Conformal prediction (CP) is a statistical framework that provides uncertainty estimation with strong mathematical guarantees [6]. This offers a promising avenue for enhancing the reliability of prostate cancer segmentation and in turn, risk factors related to it that influence the diagnosis of prostate cancer.

In this work, the primary goals include assessing the performance of conformal classification for prostate segmentation by using different confidence thresholds with internal and external cohorts.

1.2 Objectives

- Use predicted probabilities from deep learning models to quantify the uncertainty of the segmentation with a technique that offers rigorous statistical guarantees.
- Use the conformal classifier to acquire and calculate various segmentation metrics to be used in statistical analysis of the models' performance and give an accurate representation of their reliability.
- Test with different alpha thresholds (certainty) to explore their effect in the uncertainty quantification.
- To test the effect of the uncertainty quantification technique with internal and external cohorts.

1.3 Related work

In the field of prostate cancer diagnosis using MRI and deep learning (DL), automated segmentation techniques have become increasingly sophisticated. On the model Fernandez-Quilez et al. [7] developed, they demonstrated a significance decrease in DSC when tested on datasets from other institutions. Our study is closely aligned with the research conducted in this article, as we utilize datasets they acquired, we aim to help improve on the model by providing insight where we hope the quantification of uncertainty obtained will be of use. This direct connection underscores the relevance of their findings to our work in prostate cancer segmentation using MRI and deep learning.

Another emerging area in medical imaging is the application of conformal prediction (CP) techniques. Unlike traditional methods, conformal prediction offers statistical guarantees of accuracy, which are indispensable in clinical settings. The incorporation of conformal prediction to assess uncertainties in AI models has been explored notably by Olsson et al. [6]. To evaluate the CP framework they assessed efficiency, defined as the fraction of all predictions resulting in a correct single-label prediction, as well as validity or error rate. Their research underscores the potential of conformal prediction in enhancing the reliability of AI-assisted pathology, thus offering crucial insights for our study.

The quantification of uncertainty in deep learning predictions has become a focal point in enhancing clinical decision-making processes. Traditional techniques like Monte Carlo (MC) dropout and model ensembles have been employed by Gal and Gahramani [8], and Lakshminarayanan et al [9], respectively. These studies offers approximate confidence levels, however, they fall short when compared to the exact confidence levels provided by conformal prediction methods, which is crucial in high-precision medical applications. For quantification, Karimi and Samavi [5], proposes a probabilistic approach for quantifying predictive uncertainty in deep neural networks. They do this by utilizing the probabilistic existence of true labels to provide certified uncertainty bounds.

The aim of these research endeavors, including ours, is to refine the diagnostic processes for prostate cancer. Our work differs from the aforementioned works in that we use a class-conditional conformal classification technique, as well as using DSC and RVD to evaluate our models. These are noteworthy for their ability to assess the uncertainty of segmentation predictions. This provides a more robust framework for evaluating the reliability of automated segmentation models, making them increasingly relevant in medical diagnostics.

Chapter 2

Background

2.1 Clinical

2.1.1 Prostate anatomy

The prostate is an integral component of the male reproductive system. It is positioned beneath the bladder, characterized by a relatively modest size in comparison to other organs [10]. As depicted in 2.1, this anatomical structure is categorized into four distinct zones: the transition zone, central zone, anterior fibromuscular stroma, and peripheral zone. Additionally, the prostate is partitioned into three segments from top to bottom, identified as the base, mid, and apex.

Within the broader context of global health, prostate cancer (PCa) emerged as the fifth most frequently diagnosed cancer [1]. Its malignancy is on par with lung cancer among males. Significantly, 70-75% of PCa cases originate in the peripheral zone, with an additional 25% found in the transition zone, designating these regions as particularly prone to cancer development. The remaining two prostate zones rarely serve as primary sites for cancer initiation [10].

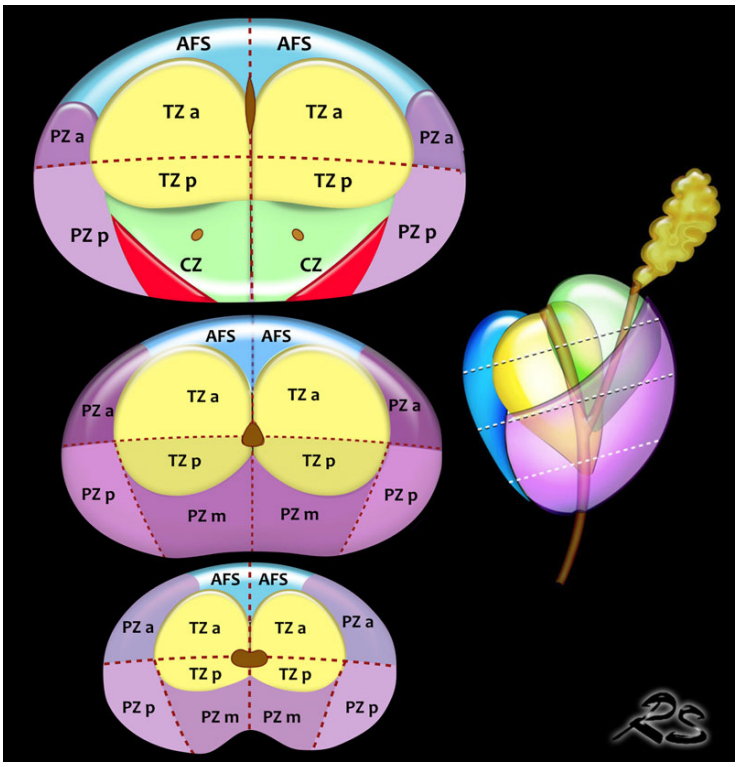

Figure 2.1: Figure of the prostate anatomy [11]. The colors yellow, green, blue, and purple represents the transition, central, anterior fibromuscular stroma, and peripheral zone, respectively. Starting from the top is the base, mid, and apex partitions.

Volume and segmentation

The volume of the prostate is a vital parameter in the diagnosis and management of prostate conditions, particularly PCa [12]. The Prostate-Specific Antigen Density (PSAd) is a crucial diagnostic metric, calculated as the Prostate-Specific Antigen (PSA) level divided by the prostate volume [7]. A higher PSAd can indicate a higher likelihood of prostate cancer. This makes accurate volume measurement essential, so that patients can refrain from undergoing unnecessary biopsies and decrease the false positives in PCa detection.

Accurate segmentation of the prostate in medical imaging is critical for determining its volume [13]. Precise segmentation allows for better assessment of prostate size and shape, which is essential for calculating PSAd. Moreover, understanding the zonal anatomy of the prostate through segmentation helps in localizing and assessing the extent of cancer in patients. Accurate prostate volume measurement and segmentation are also essential for guided biopsies and targeted therapies [14]. Guided biopsies rely heavily on precise prostate segmentation to accurately identify and sample the regions of interest. Similarly, in radiation therapy and other treatments, accurate segmentation ensures that therapy is accurately directed at cancerous tissues while minimizing impact on healthy areas [15].

These applications further underscore the importance of precise and reliable segmentation in both diagnostic and therapeutic contexts in prostate cancer management. Therefore, segmentation accuracy directly impacts the reliability of prostate volume measurements and, by extension, the effectiveness of PCa diagnostics.

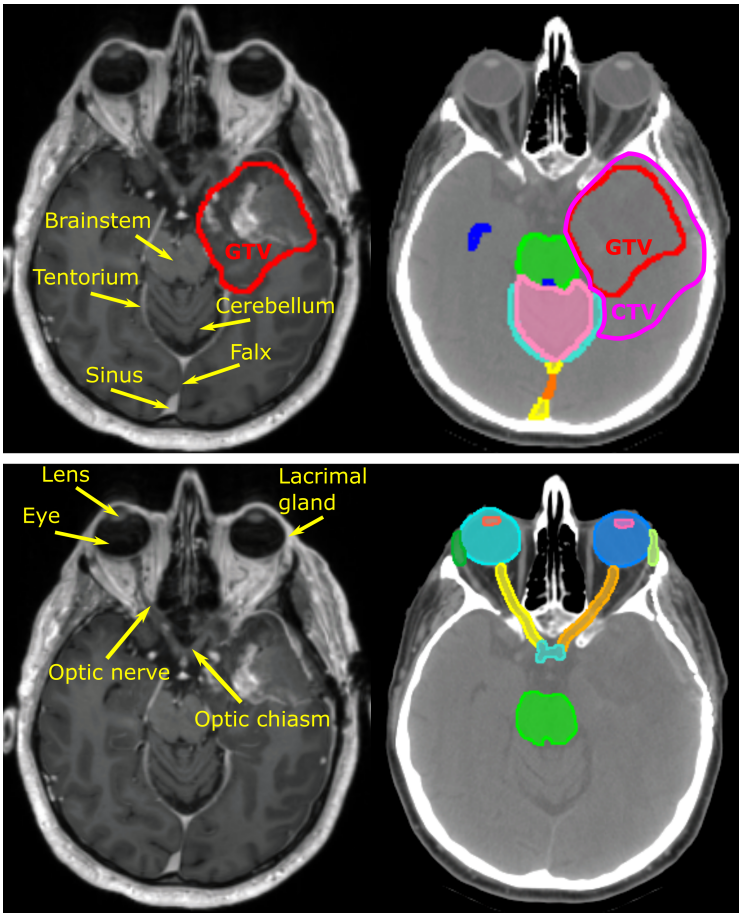

Figure 2.2: Figure of glioma image segmentations in radiation therapy [16]. Here we can observe the importance of accurate segmentations, so that healthy areas remain unharmed.

2.1.2 Magnetic resonance imaging

Magnetic Resonance Imaging (MRI) serves as a non-invasive imaging technology that exploits the interactions between strong magnetic fields and protons within the body [17]. Its optimal performance occurs in tissues characterized by high water or fat content, owing to the abundant presence of hydrogen (protons) in water and fat molecules. The choice of MRI sequence significantly influences the visual characteristics of the generated images. A sequence encompasses various radio-frequency pulses and gradients applied during the scanning process. Commonly utilized MRI sequences include T1-weighted (T1W), T2-weighted (T2W), and diffusion-weighted imaging (DWI) [18]. These sequences play a

crucial role in delineating different tissue characteristics and enhancing the diagnostic capabilities of MRI.

T2-weighted (T2W) imaging in MRI is particularly advantageous for prostate segmentation due to its high resolution and clarity in depicting prostate anatomy [3, 18]. T2W sequences provide excellent contrast between different tissue types within the prostate, making it easier to differentiate the prostate gland from surrounding tissues. This clarity is crucial for accurately delineating the prostate's boundaries, which is essential for precise volume measurement and effective segmentation. The superior image quality in T2W MRI ensures more reliable and accurate identification of prostate zones and potential lesions, enhancing the overall effectiveness of prostate cancer diagnosis and treatment planning.

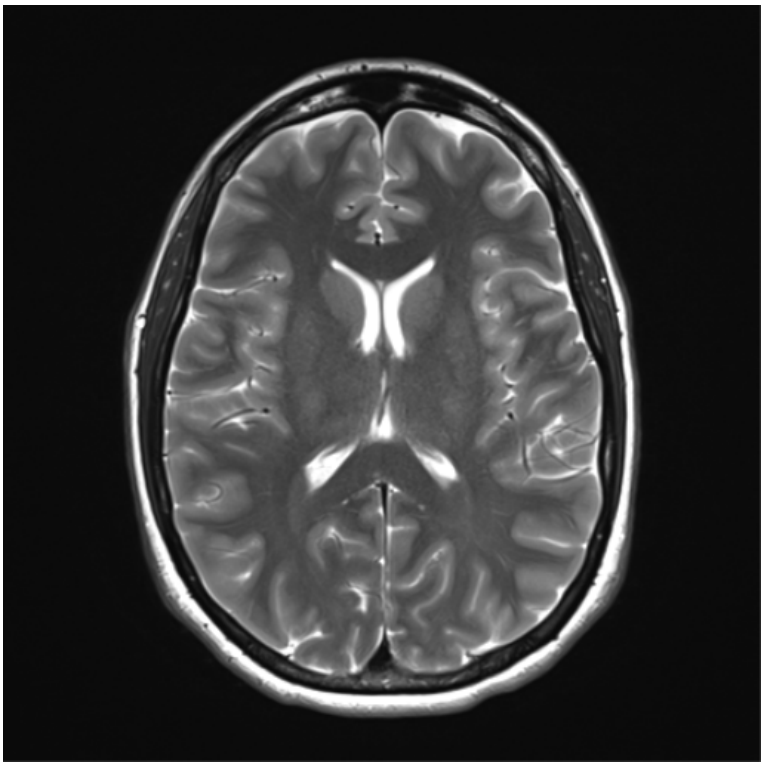

Figure 2.3: Example of a T2W MRI of the brain [19].

2.2 Technical

2.2.1 Image processing techniques

Automatic prostate segmentation

In the area of medical imaging, deep learning (DL) models has provided a sophisticated means of calculating predicted probabilities using a variety of algorithms [20]. These probabilities serve as crucial indicators in the evaluation and diagnosis of PCa.

nnU-net

The datasets used in this study was acquired by DL models using nnU-Net, which is an adaptive framework for medical image segmentation [21]. The architecture of nnU-Net is dynamically configured based on the properties of the input dataset, including decisions on using 2D, 3D, or a cascade of 2D and 3D networks. This adaptability extends to aspects like patch size, batch size, and other network hyperparameters. The nnU-Net framework automatically handles preprocessing steps such as resampling and normalization, and employs data augmentation techniques to enhance the model's robustness against variations in imaging data [7].

The nnU-Net framework is designed to output probabilistic segmentation maps. Each pixel (or voxel in 3D) in the output map represents the probability of that pixel belonging to a particular segmentation class. The network architecture is trained to predict these probabilities based on the input image data, providing us with a probability map indicating the likelihood of each pixel belonging to regions like the prostate gland, while differentiating it from surrounding tissues. This probabilistic approach enables a more refined segmentation, which is crucial for accurate medical diagnoses and analyses.

The standard U-Net architecture used in nnU-Net is well-suited for biomedical image segmentation, offering an efficient way to capture both local and global contexts from limited data [7]. The model's training process utilizes a combination of dice loss and cross-entropy with a Stochastic Gradient Descent (SGD) optimizer, and it is trained for a defined number of epochs, with the final model being an ensemble of models obtained from a cross-validation process. This setup highlights nnU-Net's capability in handling diverse datasets and its utility in developing accurate and generalizable segmentation models, making it an ideal choice for our prostate segmentation study.

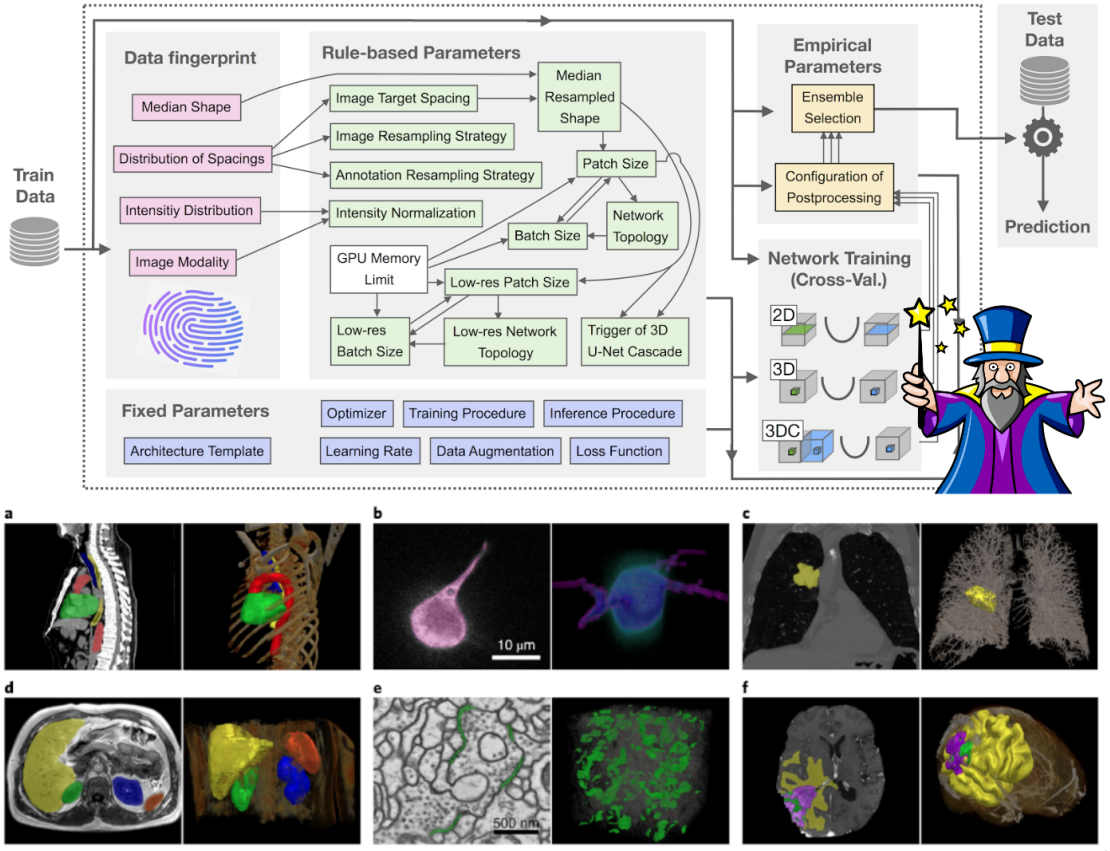

Figure 2.4: Overview of the nnU-Net framework [22].

Nearest-neighbor interpolation

In the context of image processing, nearest-neighbor interpolation emerges as a fundamental technique employed to ensure uniformity in resizing [23]. This method, applied during the image preprocessing stage, determines the value of each pixel in the resized image based on the value of the nearest pixel in the original image. Nearest-neighbor interpolation preserves the integrity of pixel values, crucial for maintaining the diagnostic accuracy of the images.

2.2.2 Uncertainty

In PCa diagnosis using MRI and DL, multiple factors can introduce uncertainty in the predictions. In MRI, uncertainty can arise from image quality, variations in anatomical structures, and the inherent complexity of interpreting soft tissue contrasts [5, 6]. DL models, while offering advanced diagnostic capabilities, also contribute to uncertainty due to the interpretive nature of model outputs [20]. This is particularly relevant in the probability-based outputs of segmentation models like nnU-Net, where the confidence of

the model in its predictions is crucial [21]. Techniques like Monte Carlo (MC) dropout and model ensembles are used to estimate uncertainty [8, 9]. Monte Carlo dropout randomly omits network units during inference to create output distributions, while model ensembles aggregate predictions from multiple models to enhance stability and reduce overfitting. Understanding and quantifying this uncertainty is vital for enhancing the reliability and interpretability of automated diagnoses, aiding clinicians in making informed decisions and planning effective treatments.

2.2.3 Conformal classification

Conformal prediction (CP), or conformal classification (CC), was utilized as the primary method employed in this study for quantifying the uncertainties of the DL models using their predicted probabilities [6]. This approach extends traditional classification methodologies by incorporating non-conformity scores, offering a deeper understanding of the uncertainty associated with each prediction. Alpha levels are key components in CC and act as confidence thresholds for determining the strictness of prediction inclusion in the conformal sets. Ranging from 0 to 1, these levels allow for a tunable balance between prediction certainty and inclusivity, enhancing the reliability and interpretability of the diagnostic process.

Class-conditional

Class-conditional conformal classifiers represent a specialized subset of Mondrian conformal classifiers, where the categorization is driven by class labels [24]. This approach is particularly pertinent in scenarios where the class of a test object is unknown, necessitating a unique treatment for these objects. In class-conditional conformal classification, a non-conformity score, or p-value, is generated for each possible class label of a test object. This score quantifies the degree of deviation or non-conformity of a test object from a given class, thereby facilitating a probabilistic assessment of class membership.

2.2.4 Metrics

Dice similarity coefficient

The Dice similarity coefficient (DSC) serves as a quantitative metric to assess the agreement between predicted and ground truth segmentation. This coefficient, ranging from 0 to 1, measures the spatial overlap between two segmentation masks, where 0 means no overlap and 1 means perfect overlap. It is calculated as the absolute difference in

total volume between two segmented regions (e.g., a ground truth model and a predicted model) relative to the volume of a reference segmentation, typically the ground truth. The DSC is defined as:

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{2.1}$$

where X and Y represent the pixel sets of the ground truth and predicted segmentations, respectively. In the context of prostate cancer diagnosis, the DSC provides valuable insights into the accuracy of the segmentation process [25, 26].

Relative volume difference

Relative Volume Difference (RVD) is a key metric in medical image analysis for quantifying the volumetric discrepancy between two segmentations [26]. It is calculated as the absolute difference in total volume between two segmented regions (e.g., a ground truth mask and a predicted mask) relative to the volume of a reference segmentation, typically the ground truth. The formula is:

$$RVD = \frac{|V_{ref} - V_{pred}|}{V_{ref}} \times 100\% \tag{2.2}$$

where V_{ref} is the reference volume and V_{pred} is the predicted volume. RVD provides a percentage indicating the extent of overestimation or underestimation by the predictive model, making it a crucial tool for evaluating the accuracy of segmentation techniques in medical imaging.

Chapter 3

Data and methods

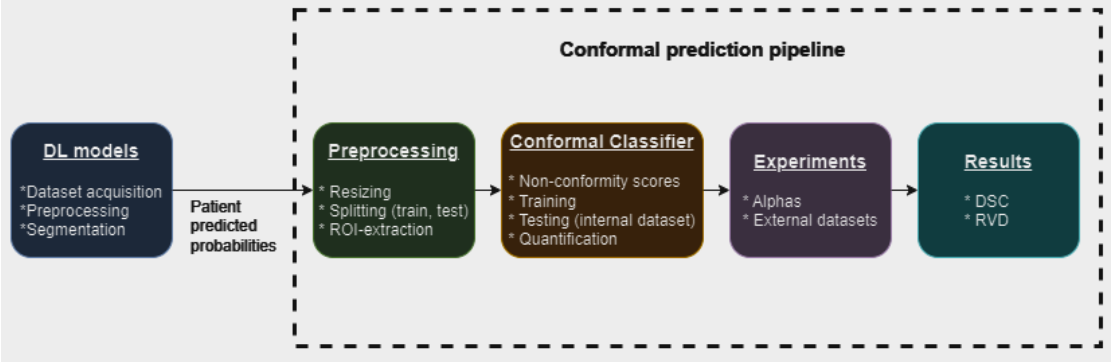

Figure 3.1: Overview of the methodology used in this work.

3.1 Setup

The methodology employed in our study necessitated the use of several specialized tools, each chosen for its specific capabilities in handling various aspects of the data processing and analysis pipeline.

- Python: Serving as the primary programming language, Python was selected for its extensive library support and its widespread use in scientific computing and machine learning. Python's versatility and readability make it ideal for implementing complex algorithms and handling large datasets.
- Crepes: For the conformal classification package, crepes was utilized for its robust capabilities in managing the uncertainty in classification tasks. This package was chosen for its simplicity in implementing class-conditional conformal classifiers, allowing us to calculate non-conformity scores for each potential class label in the MRI scans, and generating prediction sets using the classifier [24].

- SimpleITK: This package played a crucial role in our pipeline, specifically for reading MRI images. It is a simplified, open-source interface to the Insight Segmentation and Registration Toolkit (ITK), known for its efficacy in processing and analyzing high-dimensional medical images [27]. SimpleITK's functionality was essential for handling the intricacies of MRI data.
- Matplotlib: For visualization purposes, Matplotlib, a comprehensive library for creating static, animated, and interactive visualizations in Python, was employed [28]. It was particularly useful for plotting the MRI's, segmentation results, and predictions, thereby providing an intuitive understanding of the model's performance and the data characteristics.

3.2 Data

The foundation of our study rests on a meticulous selection of datasets, each contributing unique perspectives to the evaluation of our conformal classification approach. More specifically, we are using the predicted probabilities of the segmentations from the trained DL models performed by Fernandez-Quilez et al. [7], as well as the ground truths manually annotated by board-certified and approved radiologists. ProstateX, a comprehensive repository of multi-parametric MRI data, served as the primary training dataset for our conformal classifier (CC). This dataset was split into a training and testing set, and to ensure a more robust assessment, we further tested the classifier on two additional external datasets, SUS and Prostate158 (P158) [29]. This approach expanded the diversity of our study and were used for validating the generalizability of the model across distinct clinical contexts.

3.2.1 Dataset acquisition

The dataset used in this study were acquired from multiple institutions with varying technical specifications and annotation protocols [7]. Segmentations were obtained using ITKSnap v3.8 for all three datasets, and the images used are T2w images of the axial plane. The datasets are as follows:

- ProstateX: This dataset forms the core of our model training and development and consisted of 41 patients. It is a widely used open-source, single-institution dataset in prostate whole gland (WG) segmentation literature. For this dataset, the ground truth's annotation were manually conducted by two radiologists in-training and two experienced board-certified radiologists. The sequences used in the development of

the DL model were acquired with a 3.0-Tesla Siemens scanner with an in-plane resolution of 0.5 mm x 0.5 mm, 3.6 mm slice thickness and with a surface coil.

- SUS: Acquired with an in-plane resolution of 0.5 mm x 0.5 mm and a slice thickness of 3.0 mm. The annotations of the ground truths for this dataset were performed by a radiologist in-training, with the ethical approval obtained from the Regional Committee for Medical and Health Research Ethics (REC Central Norway). This dataset comprised of 48 patients and were further filtered to 41, where only the patients that were diagnosed with PCa were used in the testing.
- P158: Characterized by segmentations conducted by two board-certified radiologists, with technical acquisition parameters differing from ProstateX and SUS. This dataset consisted of 139 patients and were filtered, similarly to the SUS-dataset, down to 83 patients.

Internal and external

The internal dataset refers to the dataset used during the development and initial testing of the model, which for our case is ProstateX. It is split into training and internal testing sets. The model is trained on the training set and initially evaluated on the internal test set. External datasets are also used for testing the model's performance after it has been trained. For our study, P158 and SUS was obtained. These were served to assess the model's generalizability and robustness to new, previously unseen data.

DL models details

The preprocessing of these datasets for the DL models involved normalization of pixel intensity ranges, center cropping, axis re-ordering, and resampling of sequences to mitigate inter-site differences arising from varying acquisition protocols employed by different centers. This preprocessing was an integral part of the nn-UNet pipeline used for model development.

For segmentation, a standard nn-UNet model was employed due to its wide adoption in prostate segmentation challenges. The model was trained using the default network configuration, combining dice loss and cross-entropy with an SGD optimizer, across 1000 epochs. Data augmentation techniques were applied on the fly. A 5-fold cross-validation strategy was used, and the final model was an ensemble of the five models obtained from each cross-validation iteration. The implementation was done using TensorFlow and Python, and training and evaluation were performed on an NVIDIA A100-80G GPU.

3.3 Method

3.3.1 Pre-Processing

The pre-processing pipeline played a pivotal role in preparing the data for model training and evaluation.

Resizing

Some of the ground truth masks (GTM) had resolutions that differed from those of the predicted probabilities. To address this, we implemented a resizing technique, employing the Nearest-Neighbor interpolation method, which ensured uniformity in model training and consistency in evaluation across all datasets. This method maintains the original pixel values without introducing weighted averages or smoothing, and was chosen for its simplicity and efficiency. The Nearest-Neighbor interpolation ensures that the resized images retain the essential details crucial for accurate segmentation, contributing to the robustness of our CC.

Training, testing, and validation

We adopted a patient-centric data splitting strategy, allocating 70% of the patients from ProstateX to the training set and reserving 30% for internal testing. This partitioning comprised of 28 patients in the training set and 13 in the test set. As this work is a continuation of the study performed by Fernandez-Quilez et al. [7], using their test set of 41 patients, this partitioning was deemed as a balanced representation of the study population. The CC was then tested on the internal and external datasets, where the external datasets retained their original population.

Region of interest extraction

Due to the nature of the datasets, the training data became very large and training of the CC would at times crash the program. To counteract this we implemented a cost-cutting strategy to reduce the size of the training data. Region of interest (ROI) extraction constitutes a crucial step in focusing the analysis on clinically relevant areas and reducing overhead. In this study, ROI extraction involves capturing a specific region surrounding the prostate, incorporating a margin to encompass contextual information. This targeted approach enhances the training of the CC and quantification of the DL models uncertainty, contributing to more accurate and clinically meaningful results.

To focus the CC on clinically relevant regions and reduce training data dimensionality, we used an algorithm to extract the ROI from each slice, incorporating a 20-pixel margin around the prostate. This margin was deemed sufficient to capture contextual information relevant to prostate cancer diagnosis while minimizing unnecessary computational overhead.

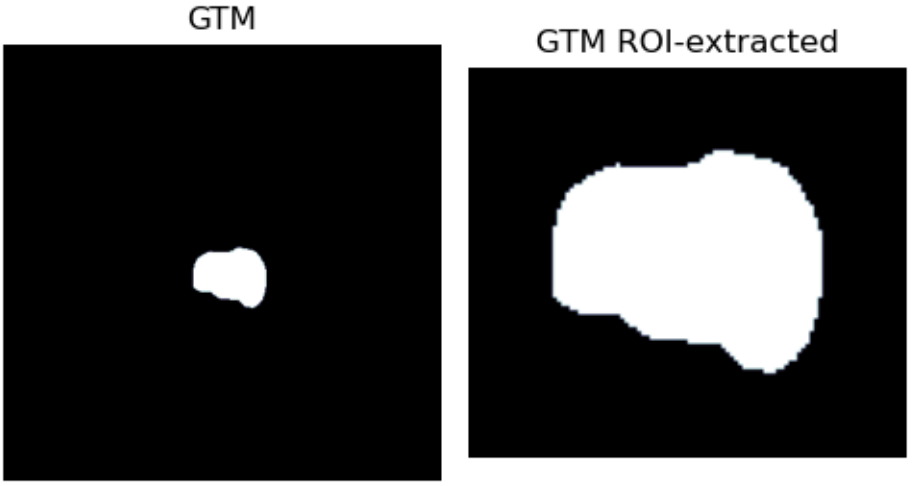

Figure 3.2: ROI-extraction with a 20 pixel margin for one slice of a GTM.

3.3.2 Class-conditional conformal classifier

For our study, we implemented a simple class-conditional conformal classifier using the crepes package [24]. Our primary methodology involves class-conditional conformal classification, an extension of traditional classification approaches that provides prediction uncertainty estimates. The CC was trained and internally tested on ProstateX. The methodology was further validated using external data from SUS and P158 to assess its generalizability and performance across diverse datasets. The conformal classification framework allows us to quantify the uncertainty associated with each prediction, providing a deeper understanding of the model's confidence. By incorporating conformal classification, we aim to enhance the interpretability and reliability of our segmentation results.

Non-conformity scores

Non-conformity scores form the basis of class conditional conformal classification. These scores quantify the degree of deviation of an individual prediction from the established class, providing a measure of uncertainty and facilitating more informed decision-making. The above results in that alphas is assigned a vector of the same length as X_prob with

a non-conformity score for each object, here defined as 1 minus the predicted probability for the correct class label:

$$\alpha = 1 - P(y|x) \tag{3.1}$$

In our implementation we utilized the **hinge** function imported from **crepes.extra** to calculate the non-conformity scores. This function takes the predicted probabilities, class names and labels as input, to create a vector of same length as X_{prob} containing the non-conformity score of each object. These scores are then used when fitting the CC. Non-conformity scores for objects where y is not known, can be acquired by using the same function using only predicted probabilities as input. This gave us an array of the same shape as the predicted probabilities with non-conformity scores for each class in the columns for each object.

Prediction sets

The output of our conformal classifier are in the form of prediction sets. Where the prediction set is in the form $Y \times 2$, where Y is the number of predictions or in our case pixels. These prediction sets can be categorized as single, multiple, or empty predictions:

- Single predictions: These occur when the CC is confident enough to assign only one class label to a test instance. This indicates a high level of certainty in the prediction.
- Multiple predictions: This happens when the classifier identifies more than one class label as a possible output for a given instance, reflecting uncertainty in the prediction. It signifies that the model cannot definitively distinguish between several class labels at the given confidence level.
- Empty predictions: These are cases where the classifier is unable to assign any class label to a test instance, typically suggesting that the instance is significantly different from the data on which the model was trained. This indicates extreme uncertainty or a lack of confidence in making any prediction.

Quantifying the total amount of these types of predictions provided us with insight into the classifier's certainty and reliability across different instances and scenarios. For simplicity, one of the methods we implemented was to convert these prediction sets to a vector of length Y, and only assigned the value 1 for the single predictions of 1, and setting the rest to 0. Furthermore, we converted this vector back to the original format e.g., (20, 400, 400), representing the number of slices, x-pixels, and y-pixels, respectively.

With this we could plot the slices where the classifier was confident in the pixels that were 1, as can be seen later in 4.1. Another method we tested was to give different labels to the different cases, so that we could visualize the uncertainty in a plot, as seen in 4.1.

Alphas (Confidence)

To quantify the impact of uncertainty and alpha levels on our model's performance, we tested the CC on a range of alpha values, including 0.25, 0.50, 0.75, and 0.95. These alpha levels correspond to different confidence thresholds, allowing us to assess the classifier's performance using various degrees of certainty. The lower alpha values (e.g., 0.25) represent a higher confidence in the predictions, whereas higher alpha values (e.g., 0.95) indicate greater caution and a lower threshold for accepting predictions as certain.

3.3.3 Uncertainty assessment

Validity

Validity is a measure of how often the model's prediction sets correctly include the true label. It assesses the reliability of the model in providing prediction sets that encompass the actual outcomes [30]. In this work, the validity is calculated as:

$$Validity = \frac{Correct\_classifications + Multiple\_predictions}{Total\_predictions} \tag{3.2}$$

Where **Correct_classifications** are instances where the prediction set contains only one label, which matches the true label, and **Multiple_predictions** refer to instances where the prediction set includes more than one label, thereby covering the true label. This measure of validity ensures that the prediction sets accurately reflect the true outcomes in a significant proportion of cases.

Efficiency

Efficiency is a measure of how often the model makes specific and informative predictions. It evaluates the model's ability to provide prediction sets that are not only correct but also precise, typically represented by correct single-label predictions [6]. We have calculated efficiency as the fraction of correct single predictions:

$$Efficiency = \frac{Correct\_classifications}{Total\_predictions} \tag{3.3}$$

This measure of efficiency reflects the model's precision and its effectiveness in making definitive predictions.

3.3.4 Segmentation assessment

DSC

The DSC was computed for each slice of each patient, only taking into account the slices that contained a mask. For each patient, we calculated the mean DSC across these slices, and then the overall mean DSC was computed across all patients. Additionally, we stratified the analysis by dividing the slices into base, mid, and apex sections to assess regional differences in segmentation accuracy. DSC calculations were performed to compare:

- GTM and CC segmentation.
- GTM and predicted probabilities, applying a threshold of 0.5, where values of 0.5 or higher were classified as 1, and values below 0.5 as 0.

This approach allowed us to evaluate the segmentation performance from multiple perspectives, ensuring a thorough understanding of our model's accuracy.

RVD

For RVD, the volume was determined by calculating the volumes across all slices for each segmentation. This approach provided us with a singular, overall measurement of the volume difference, reflecting the total discrepancy in volume between two segmentations across the entire scanned volume. Such an approach is crucial in clinical and research contexts, particularly for evaluating the overall size of a tumor or an organ. This singular measure of volume difference offered us insights into the macroscopic accuracy of the segmentation, essential for clinical assessments and comparative studies.

Chapter 4

Results

4.1 Methodology

The code will not be included as the work is under consideration for publication.

4.2 Experimental Results

ProstateX results

AI points predictions	
Correct predictions	36498481
Incorrect predictions	27080
Accuracy (%)	99.93%
Error (%)	0.07%
DSC All	0.9834 ±0.0019
DSC Base	0.9783 ±0.0044
DSC Mid	0.9863 ±0.002
DSC Apex	0.9833 ±0.0014
RVD (%)	0.0002%
Total Volume PP	942003
Total Volume GTM	942001

Table 4.1: Metrics acquired from the predicted probabilities (PP) of ProstateX using a threshold on 0.5. DSC and RVD are compared against the ground truth masks (GTM) of ProstateX.

Conformal classifier				
Prediction sets	Confidence level (%)			
	25%	50%	75%	95%
Empty	677982	439374	242078	133943
Single	35847579	36086187	36283483	36391618
Multiple	0	0	0	0
Applied statistics				
DSC All	0.5986 ±0.2451	0.7728 ±0.1726	0.8758 ±0.0723	0.9729 ±0.0052
DSC Base	0.7023 ±0.278	0.7731 ±0.176	0.7908 ±0.2523	0.9689 ±0.0071
DSC Mid	0.5188 ±0.4024	0.7731 ±0.3343	0.9582 ±0.0114	0.9772 ±0.008
DSC Apex	0.6274 ±0.2525	0.7731 ±0.176	0.8566 ±0.0764	0.9711 ±0.0071
RVD PP (%)	59.02%	33.69%	12.74%	4.02%
RVD GTM (%)	59.02%	33.69%	12.74%	4.01%
Total Volume	386056	624664	821960	904181
Uncertainty				
Correct single predictions	35847531	36085921	36282748	36386168
Points that should be 1	555971	317581	120754	42713
Points that should be 0	122059	122059	122059	96680
Efficiency (%)	98.14%	98.80%	99.34%	99.62%
Validity (%)	98.14%	98.80%	99.34%	99.62%

Table 4.2: ProstateX results

SUS

AI points predictions	
Correct predictions	211495460
Incorrect predictions	24540
Accuracy (%)	99.9%
Error (%)	0.01%
DSC All	0.9648 ±0.0672
DSC Base	0.9999 ±0.0002
DSC Mid	0.9891 ±0.0363
DSC Apex	0.7853 ±0.3584
RVD (%)	0.6958%
Total Volume PP	3550953
Total Volume GTM	3526413

Table 4.3: Metrics acquired from the predicted probabilities (PP) of SUS using a threshold on 0.5. DSC and RVD are compared against the ground truth masks (GTM) of SUS.

Conformal classifier				
Prediction sets	Confidence level (%)			
	25%	50%	75%	95%
Empty	3734128	3064130	2643608	1394302
Single	207785872	208455870	208876392	210125698
Multiple	0	0	0	0
Applied statistics				
DSC All	0.5336 ±0.1963	0.6659 ±0.1619	0.7263 ±0.1269	0.8716 ±0.0771
DSC Base	0.5303 ±0.2497	0.5913 ±0.2451	0.6226 ±0.2472	0.7858 ±0.2487
DSC Mid	0.4895 ±0.3315	0.6737 ±0.2928	0.7665 ±0.2232	0.9251 ±0.0771
DSC Apex	0.4255 ±0.248	0.4908 ±0.2741	0.5235 ±0.2885	0.6843 ±0.3083
RVD PP (%)	64.92%	46.05%	34.21%	8.33%
RVD GTM (%)	64.67%	45.67%	33.75%	7.69%
Total Volume	1245761	1915759	2336281	3255258
Uncertainty				
Correct single predictions	207784099	208453635	208873853	210118090
Points that should be 1	2282425	1612889	1192671	278763
Points that should be 0	1453476	1453476	1453476	1123147
Efficiency (%)	98.23%	98.55%	98.75%	99.34%
Validity (%)	98.23%	98.55%	98.75%	99.34%

Table 4.4: SUS results

P158

AI points predictions	
Correct predictions	196921390
Incorrect predictions	27834
Accuracy (%)	99.9%
Error (%)	0.01%
DSC All	0.9515 ±0.1251
DSC Base	0.9525 ±0.1653
DSC Mid	0.9722 ±0.1083
DSC Apex	0.7852 ±0.378
RVD (%)	0.3814%
Total Volume PP	7324819
Total Volume GTM	7296985

Table 4.5: Metrics acquired from the predicted probabilities (PP) of P158 using a threshold on 0.5. DSC and RVD are compared against the ground truth masks (GTM) of P158.

Conformal classifier				
Predictions	Confidence level (%)			
	25%	50%	75%	95%
Empty	11342642	10308883	9736253	6545683
Single	185606582	186640341	187212971	190403541
Multiple	0	0	0	0
Applied				
DSC All	0.3523 ±0.1739	0.4353 ±0.1803	0.4712 ±0.183	0.6624 ±0.1665
DSC Base	0.2443 ±0.2191	0.2854 ±0.243	0.3107 ±0.2573	0.5062 ±0.294
DSC Mid	0.3897 ±0.2248	0.4993 ±0.2251	0.5443 ±0.2249	0.7318 ±0.2002
DSC Apex	0.1809 ±0.2188	0.211 ±0.2462	0.2283 ±0.2622	0.3707 ±0.3499
RVD PP (%)	67.25%	53.13%	45.32%	23.71%
RVD GTM (%)	67.12%	52.96%	45.11%	23.42%
Total Volume	2399079	3432838	4005468	5588348
Uncertainty				
Correct single predictions	185606582	186640341	187212971	190400547
Points that should be 1	4897906	3864147	3291517	1711631
Points that should be 0	6444736	6444736	6444736	4837046
Efficiency (%)	94.24%	94.77%	95.06%	96.67%
Validity (%)	94.24%	94.77%	95.06%	96.67%

Table 4.6: P158 results.

Uncertainty plots

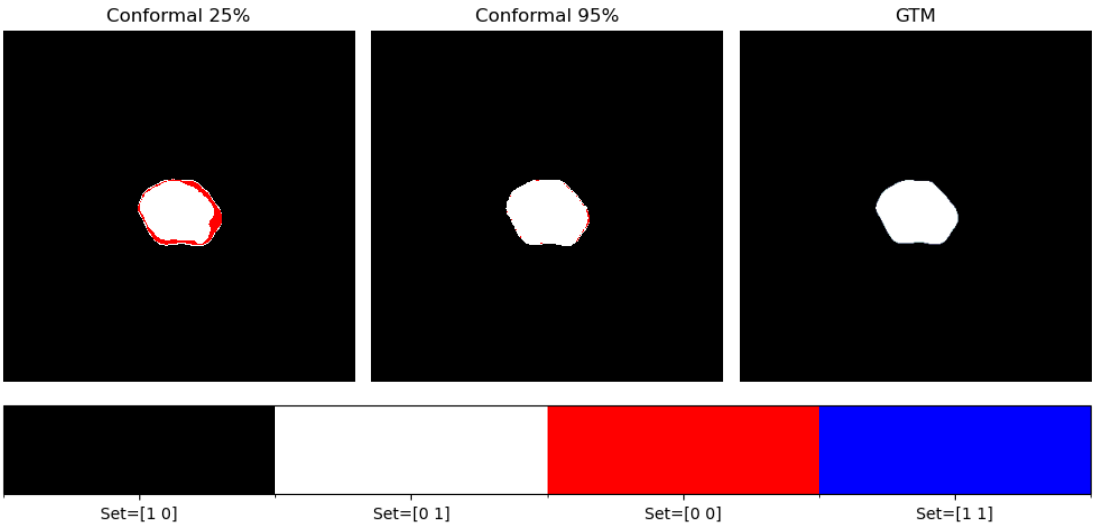

Figure 4.1: Visualization of the uncertainty between two different alpha levels, including the GTM.

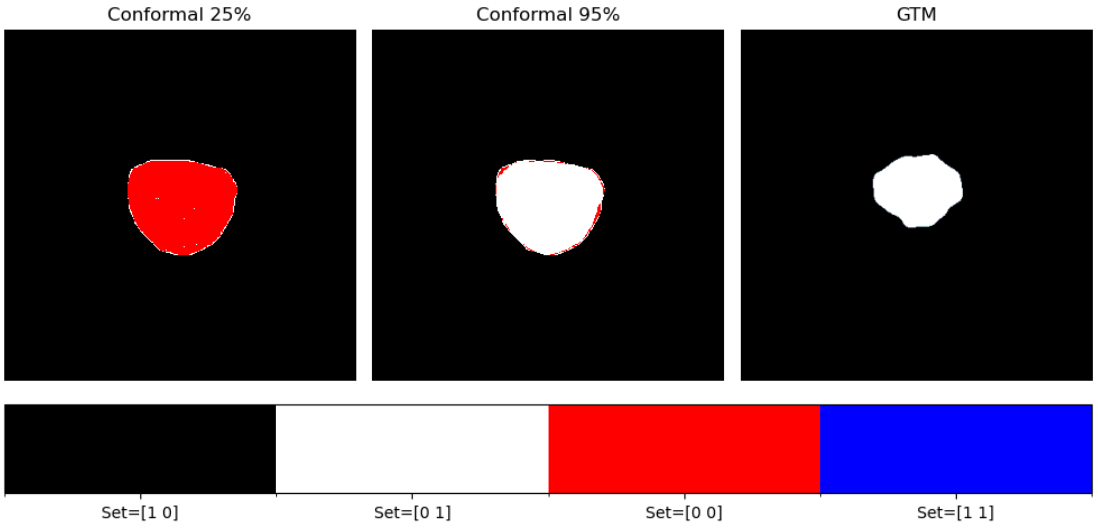

Figure 4.2: Visualization of the uncertainty between two different alpha levels, including the GTM.

4.2.1 Observations

An analysis of the DSC and RVD at different alpha levels reveals a distinct trend. At a 25% alpha level, the DSC score is observed to be lower compared to the scores as the alpha levels increases, as depicted in e.g., table 4.2. The RVD starts out higher and decreases as the alpha level increases. A notable aspect of this trend is the high number of empty sets at the lower alpha levels. This suggests a direct relationship between the alpha level and the model's confidence in making predictions, where a lower alpha level leads to a higher frequency of instances in which the model does not make any prediction, due to the stricter confidence threshold. From the uncertainty plots 4.1, and 4.2, we can observe two extremely different instances where the lower alpha levels greatly impact the uncertainty of the predictions. We observe no multi-label predictions. For almost all the metrics, we can see a consistent increase or decrease as the alpha levels change, with lower levels performing the worst.

When calculating for the uncertainty assessment we can find high percentages of validity and efficiency, indicating that the model is not only consistently including the true label in its predictions, but is also frequently making precise and correct single-label predictions. This combination suggests a well-calibrated and reliable model that is adept at making predictions with a high degree of certainty, offering valuable insights for the use of conformal prediction in prostate segmentation.

Chapter 5

Discussion

Conformal prediction vs MC dropout and model ensembles

In the context of PCa diagnosis using MRI and DL, Monte Carlo (MC) dropout and model ensembles are notable techniques for handling uncertainty. MC dropout functions by randomly dropping units in neural networks during inference to generate a distribution of outputs, thereby estimating uncertainty. Model ensembles use multiple models to make predictions, averaging their outputs to improve reliability and reduce overfitting. However, these methods only provide approximations of confidence levels and lack the strong statistical guarantees offered by conformal methods. Conformal prediction, unlike MC dropout or ensembles, can provide exact confidence levels (e.g., 95%), ensuring more precise and statistically robust uncertainty quantification. This exactness in confidence levels is crucial for clinical decision-making and is a primary reason for choosing conformal methods over traditional DL techniques in this study.

Impact of alpha levels on metrics

We could observe a recurring trend in the quantified metrics across the different alpha levels, which can be explained by the underlying mechanics of conformal prediction and its reliance on confidence thresholds. At a lower alpha level, the model has a stricter criterion for prediction inclusion, resulting in a higher number of empty sets. These empty sets indicate instances where the model's predictions do not meet the confidence threshold. As the alpha level gradually increases, the model's confidence threshold becomes less strict, allowing for more predictions to be made, including those that might be less certain. This decrease in empty sets at higher alpha levels likely contributes to the higher DSC's and lower RVD's observed, as more predictions are made and evaluated against the true labels. The increase or decrease in the metrics at higher alpha levels,

therefore, reflects the balance between the model's confidence in its predictions and the inclusivity of potential outcomes within the prediction sets. The absence of multi-label predictions in our study likely stems from the high calibration quality of the DL model we extended from the previous study of Fernandez-Quilez et al. [7]. This suggests that the model's robust training and fine-tuning have resulted in precise and confident predictions.

Our prediction set handling

In calculating the DSC and RVD, a simplified approach was employed. This involved assigning a value of 0 to every case where the prediction set was not [0 1]. The rationale behind this approach was to focus on instances where the conformal classifier was highly confident that a pixel would be classified as 1. This method prioritizes the identification of pixels confidently predicted as belonging to class 1, contributing to an uncertainty in the perceived volume of correctly classified pixels.

The primary limitation of this approach is its potential impact on the overall accuracy and reliability of the DSC score and RVD's, as a performance metric. Specifically, the approach reduces sensitivity to Class 0. Simplifying these metric's calculations to prioritize class 1 predictions, the model's ability to correctly identify class 0 is not thoroughly reflected. This could lead to an overestimation of the model's performance in scenarios where identifying class 0 is equally important. There is a potential overestimation of performance for class 1 predictions, as might be seen in figure 4.2. In this figure it can clearly be observed that the volume is notably larger than the GTM. The methodology might inflate the perceived performance for class 1 predictions. In cases where the model fails to make a prediction (empty sets), assigning a 0 value might not accurately represent the true uncertainty or potential error in classification. The impact on model evaluation is significant. This approach focuses on the model's certainty in classifying pixels as 1, potentially overlooking the broader evaluation of the model's overall predictive capabilities, including its performance in situations of uncertainty or in predicting class 0. The chosen methodology, while offering insights into the model's performance in certain aspects, might not provide the comprehensive view we are looking for in its overall predictive accuracy and reliability.

Enhancing AI reliability with prediction sets

In the study conducted by Olsson et al. [6], they proposed the idea of flagging the unreliable predictions for human intervention. The primary benefits of this approach is that it enhances patient safety by reducing AI errors and opens up the opportunity for a collaborative workflow between AI and human experts. In such a setting, the AI

would handle the straightforward cases and the experts would focus on the complex ones. This approach has been shown to significantly lower error rates compared to traditional methods. However, it also introduces potential issues, such as the risk of an unmanageably high number of unreliable predictions and the added complexity of manually reviewing these cases.

Uncertainty assessment

The results of our study prominently feature high percentages of efficiency and validity in the uncertainty assessment of prostate cancer segmentation. Efficiency percentages ranged impressively from 94.24% to 99.34% on external datasets, with an equal measure of validity. Comparing this to the study conducted by Olsson et al.'s [6] 78%, indicates that our model is not only consistent in the inclusion of true label in its predictions, but also precision in its correct single-label predictions. These tables showcases the model's exceptional calibration and reliability, and also its adeptness at making predictions with significant certainty. This level of performance is particularly critical in prostate segmentation, as it highlights the model's capacity to effectively handle uncertainty, ensuring trustworthy and precise predictions.

Chapter 6

Conclusions

This thesis presents a significant advancement in prostate cancer diagnosis through the integration of Conformal Prediction (CP) with deep learning models for prostate segmentation in MRI analysis. By utilizing CP, the study effectively quantifies uncertainties in segmentation, highlighting its robustness across varied datasets. Additionally, the impact of different alpha levels were tested to explore how varying levels of confidence thresholds influence the model's performance. This analysis provides deeper insights into the trade-offs between predictive accuracy and uncertainty, guiding the optimal configuration of the conformal prediction framework. The use of metrics like DSC and RVD, along with high efficiency and validity rates (94.24% to 99.34%), underscores the precision and reliability of the approach. This research not only enhances the accuracy of prostate cancer diagnosis but also contributes to the broader field of medical imaging, demonstrating the potential of CP in improving clinical decision-making and patient care.

6.1 Future Directions

Although the CC implemented in this study achieved high efficiency and validity scores, the DSC and RVD scores were not comparable. It would be interesting to see other approaches to handling the uncertainties in DL models as well. The final goal of the thesis was to develop an automatic quality control of prostate segmentation, however, we never advanced that far. With that in mind, some potential future direction could be:

- Expanding to automatic quality control
- Exploring alternative uncertainty handling approaches

- Methods for improving DSC and RVD scores

Bibliography

[1] Prashanth Rawla. Epidemiology of Prostate Cancer. *World Journal of Oncology*, 10(2):63–89, April 2019. ISSN 1920-4531. doi: 10.14740/wjon1191. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6497009/.

[2] Gary J. Kelloff, Peter Choyke, and Donald S. Coffey. Challenges in Clinical Prostate Cancer: Role of Imaging. *AJR. American journal of roentgenology*, 192(6):1455–1470, June 2009. ISSN 0361-803X. doi: 10.2214/AJR.09.2579. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2893141/.

[3] Sarah Montagne, Dimitri Hamzaoui, Alexandre Allera, Malek Ezziane, Anna Luzurier, Raphaelle Quint, Mehdi Kalai, Nicholas Ayache, Hervé Delingette, and Raphaële Renard-Penna. Challenge of prostate MRI segmentation on T2-weighted images: inter-observer variability and impact of prostate morphology. *Insights into Imaging*, 12(1):71, June 2021. ISSN 1869-4101. doi: 10.1186/s13244-021-01010-9. URL https://doi.org/10.1186/s13244-021-01010-9.

[4] Zhiqiang Tian, Lizhi Liu, Zhenfeng Zhang, and Baowei Fei. PSNet: prostate segmentation on MRI based on a convolutional neural network. *Journal of Medical Imaging*, 5(2):021208, January 2018. ISSN 2329-4302, 2329-4310. doi: 10.1117/1.JMI.5.2.021208. URL https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-5/issue-2/021208/PSNet--prostate-segmentation-on-MRI-based-on-a-convolutional/10.1117/1.JMI.5.2.021208.full. Publisher: SPIE.

[5] Hamed Karimi and Reza Samavi. Quantifying Deep Learning Model Uncertainty in Conformal Prediction, June 2023. URL http://arxiv.org/abs/2306.00876. arXiv:2306.00876 [cs].

[6] Henrik Olsson, Kimmo Kartasalo, Nita Mulliqi, Marco Capuccini, Pekka Ruusuvuori, Hemamali Samaratunga, Brett Delahunt, Cecilia Lindskog, Emiel A. M. Janssen, Anders Blilie, Lars Egevad, Ola Spjuth, and Martin Eklund. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature Communications*, 13:7761, December 2022. ISSN 2041-1723.

doi: 10.1038/s41467-022-34945-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9755280/.

[7] Alvaro Fernandez-Quilez, Tobias Nordström, Trygve Eftestøl, Andreas Bremset Alvestad, Fredrik Jäderling, Svein Reidar Kjosavik, and Martin Eklund. Revisiting prostate segmentation in magnetic resonance imaging (MRI): On model transferability, degradation and PI-RADS adherence. preprint, Radiology and Imaging, August 2023. URL http://medrxiv.org/lookup/doi/10.1101/2023.08.21.23294376.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR, June 2016. URL https://proceedings.mlr.press/v48/gal16.html. ISSN: 1938-7228.

[9] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

[10] Emil Obrestad. Prostate Lesion Detection on Apparent Diffusion Coefficient MRI based on Convolutional Neural Networks. Master's thesis, uis, 2021. URL https://uis.brage.unit.no/uis-xmlui/handle/11250/2788788. Accepted: 2021-10-08T15:51:18Z ISSN: 7308-5243.

[11] Rhiannon van Loenhout, Frank Zijta, Robin Smithuis, and Ivo Schoots. The Radiology Assistant : Prostate Anatomy, June 2023. URL https://radiologyassistant.nl/abdomen/prostate/prostate-cancer-pi-rads-v2-1-1.

[12] Thomas A. Stamey, Fuad S. Freiha, John E. McNeal, Elise A. Redwine, Alice S. Whittemore, and Hans-Peter Schmid. Localized prostate cancer. Relationship of tumor volume to clinical significance for treatment of prostate cancer. *Cancer*, 71(S3):933–938, 1993. ISSN 1097-0142. doi: 10.1002/1097-0142(19930201)71:3+<933::AID-CNCR2820711408>3.0.CO;2-L. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142%2819930201%2971%3A3%20%3C933%3A%3AAID-CNCR2820711408%3E3.0.CO%3B2-L. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0142%2819930201%2971%3A3%2B%3C933%3A%3AAID-CNCR2820711408%3E3.0.CO%3B2-L.

[13] Nooshin Ghavami, Yipeng Hu, Eli Gibson, Ester Bonmati, Mark Emberton, Caroline M. Moore, and Dean C. Barratt. Automatic segmentation of prostate MRI

using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. *Medical Image Analysis*, 58:101558, December 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.101558. URL https://www.sciencedirect.com/science/article/pii/S1361841519301008.

[14] Emran Mohammad Abu Anas, Parvin Mousavi, and Purang Abolmaesumi. A deep learning approach for real time prostate segmentation in freehand ultrasound guided biopsy. *Medical Image Analysis*, 48:107–116, August 2018. ISSN 1361-8415. doi: 10.1016/j.media.2018.05.010. URL https://www.sciencedirect.com/science/article/pii/S1361841518303499.

[15] Timo Kiljunen, Saad Akram, Jarkko Niemelä, Eliisa Löyttyniemi, Jan Seppälä, Janne Heikkilä, Kristiina Vuolukka, Okko-Sakari Kääriäinen, Vesa-Pekka Heikkilä, Kaisa Lehtiö, Juha Nikkinen, Eduard Gershkevitsh, Anni Borkvel, Merve Adamson, Daniil Zolotuhhin, Kati Kolk, Eric Pei Ping Pang, Jeffrey Kit Loong Tuan, Zubin Master, Melvin Lee Kiang Chua, Timo Joensuu, Juha Kononen, Mikko Myllykangas, Maigo Riener, Miia Mokka, and Jani Keyriläinen. A Deep Learning-Based Automated CT Segmentation of Prostate Cancer Anatomy for Radiation Therapy Planning-A Retrospective Multicenter Study. *Diagnostics*, 10(11):959, November 2020. ISSN 2075-4418. doi: 10.3390/diagnostics10110959. URL https://www.mdpi.com/2075-4418/10/11/959. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[16] Nadya Shusharina and Thomas Bortfeld. Glioma Image Segmentation for Radiotherapy: RT targets, barriers to cancer spread, and organs at risk (GLIS-RT), 2021. URL https://www.cancerimagingarchive.net/collection/glis-rt/.

[17] Robert-Jan M. van Geuns, Piotr A. Wielopolski, Hein G. de Bruin, Benno J. Rensing, Peter M. A. van Ooijen, Marc Hulshoff, Matthijs Oudkerk, and Pim J. de Feyter. Basic principles of magnetic resonance imaging. *Progress in Cardiovascular Diseases*, 42(2):149–156, September 1999. ISSN 0033-0620. doi: 10.1016/S0033-0620(99)70014-9. URL https://www.sciencedirect.com/science/article/pii/S0033062099700149.

[18] Hasan Aydin, Volkan Kizilgöz, Idil Günes Tatar, Çagr Damar, Ali Riza Ugan, Irem Paker, and Baki Hekimoglu. Detection of Prostate Cancer With Magnetic Resonance Imaging: Optimization of T1-Weighted, T2-Weighted, Dynamic-Enhanced T1-Weighted, Diffusion-Weighted Imaging Apparent Diffusion Coefficient Mapping Sequences and MR Spectroscopy, Correlated With Biopsy and Histopathological Findings. *Journal of Computer Assisted Tomography*, 36(1):30, February 2012. ISSN 0363-8715. doi: 10.1097/RCT.

0b013e31823f6263. URL https://journals.lww.com/jcat/abstract/2012/01000/detection_of_prostate_cancer_with_magnetic.6.aspx.

[19] Jeremy Jones. T2 weighted image | Radiology Reference Article | Radiopaedia.org. URL https://radiopaedia.org/articles/t2-weighted-image.

[20] Biraja Ghoshal, Allan Tucker, Bal Sanghera, and Wai Lup Wong. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence*, 37(2):701–734, 2021. ISSN 1467-8640. doi: 10.1111/coin.12411. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12411. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/coin.12411.

[21] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. URL https://www.nature.com/articles/s41592-020-01008-z. Number: 2 Publisher: Nature Publishing Group.

[22] Fabian Isensee. nnUNet/documentation/assets/nnU-Net_overview.png at master · MIC-DKFZ/nnUNet. URL https://github.com/MIC-DKFZ/nnUNet/blob/master/documentation/assets/nnU-Net_overview.png.

[23] Rukundo Olivier and Cao Hanqiang. Nearest Neighbor Value Interpolation. *International Journal of Advanced Computer Science and Applications*, 3(4), 2012. ISSN 2158107X, 21565570. doi: 10.14569/IJACSA.2012.030405. URL http://thesai.org/Publications/ViewPaper?Volume=3&Issue=4&Code=IJACSA&SerialNo=5.

[24] Henrik Bostrom. crepes: a Python Package for Generating Conformal Regressors and Predictive Systems.

[25] Baris Turkbey, Sergei V. Fotin, Robert J. Huang, Yin Yin, Dagane Daar, Omer Aras, Marcelino Bernardo, Brian E. Garvey, Juanita Weaver, Hrishikesh Haldankar, Naira Muradyan, Maria J. Merino, Peter A. Pinto, Senthil Periaswamy, and Peter L. Choyke. Fully Automated Prostate Segmentation on MRI: Comparison With Manual Segmentation Methods and Specimen Volumes. *American Journal of Roentgenology*, 201(5):W720–W729, November 2013. ISSN 0361-803X. doi: 10.2214/AJR.12.9712. URL https://www.ajronline.org/doi/full/10.2214/AJR.12.9712. Publisher: American Roentgen Ray Society.

[26] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias

Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip "Eddie" Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, February 2014. ISSN 1361-8415. doi: 10.1016/j.media.2013.12.002. URL https://www.sciencedirect.com/science/article/pii/S1361841513001734.

[27] Ziv Yaniv, Bradley C. Lowekamp, Hans J. Johnson, and Richard Beare. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *Journal of Digital Imaging*, 31(3):290–303, June 2018. ISSN 1618-727X. doi: 10.1007/s10278-017-0037-8. URL https://doi.org/10.1007/s10278-017-0037-8.

[28] Thomas A. Caswell, Elliott Sales de Andrade, Antony Lee, Michael Droettboom, Tim Hoffmann, Jody Klymak, John Hunter, Eric Firing, David Stansby, Nelle Varoquaux, Jens Hedegaard Nielsen, Oscar Gustafsson, Kyle Sunden, Benjamin Root, Ryan May, Phil Elson, Jouni K. Seppänen, hannah, Jae-Joon Lee, Darren Dale, Damon McDougall, Andrew Straw, Paul Hobson, Greg Lucas, Ruth Comer, Christoph Gohlke, Adrien F. Vincent, Tony S. Yu, Eric Ma, and Steven Silvester. matplotlib/matplotlib: REL: v3.7.4, November 2023. URL https://zenodo.org/records/10152802.

[29] Lisa C. Adams, Marcus R. Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M. Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K. Bressem. Prostate158 - An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148:105817, September 2022. ISSN 1879-0534. doi: 10.1016/j.compbiomed.2022.105817.

[30] Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting With Confidence: Using Conformal Prediction in Drug Discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, January 2021. ISSN 0022-3549. doi: 10.1016/j.xphs.2020.09.055. URL https://www.sciencedirect.com/science/article/pii/S002235492030589X.